Discovery and efficient reuse of technology pictures using Wikimedia infrastructures. A proposal

Lambert Heller and **Ina Blümel** and **Simone Cartellieri**
Technische Informationsbibliothek Hannover
Welfengarten 1B
D-30167 Hannover

Christian Wartena
Hochschule Hannover
Expo Plaza 12
D-30539 Hannover

Motivation

Multimedia objects, especially images and figures, are essential for the visualization and interpretation of research findings. The distribution and reuse of these scientific objects is significantly improved under open access conditions, for instance in Wikipedia articles, in research literature, as well as in education and knowledge dissemination, where licensing of images often represents a serious barrier.

Whereas scientific publications are retrievable through library portals or other online search services due to standardized indices there is no targeted retrieval and access to the accompanying images and figures yet. Consequently there is a great demand to develop standardized indexing methods for these multimedia open access objects in order to improve the accessibility to this material.

With our proposal, we hope to serve a broad audience which looks up a scientific or technical term in a web search portal first. Until now, this audience has little chance to find an openly accessible and reusable image narrowly matching their search term on first try - frustratingly so, even if there **is** in fact such an image included in some open access article.

Aim

To address this objective a process for automatic harvesting and indexing of multimedia open access objects needs to be developed. Wikimedia Commons [1], operated by the Wikimedia Foundation, is one of the most important open access media collections and a platform and service for the harvesting and distribution thereof. Wikimedia Commons collects freely licensed objects, mainly images and figures, including their metadata and provides them to Wikipedia and all other web users.

Obviously, Wikipedia itself is since many years one of the most frequent targets for web searches on all topics from the area of science and education. Wikidata is already in use for including open access literature more systematically into Wikipedia articles, since Wikidata is already used as a store of references to scholarly articles and other scholarly artifacts.

[1] https://commons.wikimedia.org/

The aim of our proposal is to develop a process for the automatic harvesting and indexing of articles including their images and figures derived from quality controlled open access journals in the fields of engineering and technology, using the infrastructure of Wikimedia Commons, Wikisource and Wikidata.

Images in scientific literature

Based on a rough estimation, we expect about 2-3 pictures (apart from graphs, formula, tables) in each journal article from technology and engineering disciplines. Usually, images contained in research papers are key to quick understanding of the content. At the same time, they are highly reusable: In education as well as in journalism, and ultimately in other research publications.

De facto, image search in research and education often starts with popular web search engine approaches, like Google image search (and similar). Google image search can be restricted to material that is licensed for reuse, but search terms are not matched to a set of standardized terms (or semantic entities). Only this so-called thesaurus based search is known to deliver high precision results in information retrieval, since many search terms are ambigous. This is even more true for information retrieval in science.

Having said this, some disciplines are highly centered on digital imagery, like biology and medicine, or geo sciences and astronomy. They are better prepared, insofar they often have specialized solutions to classify and retrieve images.[2] In broader fields, like technology and engineering, we are not aware of a thesaurus-based retrieval solution for pictures, let alone one across multiple publishers or sources, allowing for a specialized search on material that is licensed for reuse.

Method

We want to achieve a large-scale thesaurus based content mining solution (Blümel et al. 2014) for articles in the field of technology and engineering, focused on picture captions and picture references within the articles. In order to build a relevantly huge corpus that can be easily processed with textmining approaches, we focus on a few large, established

[2] e.g. SSOIS – Solar System Object Image Search http://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/ssois/

journal publishers which offer their article contents solely in XML format and under CC-BY license conditions.

Harvesting and picture extraction Among the 20 largest OASPA[3] publishers there are five (PLoS, Frontiers, Hindawi, SpringerOpen, BioMedCentral), which offer all of their article content in XML format, and additionally one publisher (MDPI), that offers most of its articles in XML, and at the time offer a bulk download option, which supports easier text processing on a large number of articles. Additionally, all of the articles fall under a CC BY license. From these six publishers, we found 90 journals that publish mostly or in parts articles from technology and engineering. (Often, these articles come from mixed fields, like e.g. biomedical engineering.) From the beginning of 2013 until August 2015, these 90 journals alone published around 35.000 articles. Therefore, we expect that this article corpus may contain roughly 100.000 images of all kinds which are available under open access conditions. All in all, this is a promising corpus, which should be processable with automatic methods relative easily.

To reach this goal it is planned to harvest scientific publications and subsequently extract and enrich the metadata of the accompanying images and figures in order to generate a standardized index. All of these steps can be done utilizing the Wikimedia projects mentioned before. (Cf. Figure 1)

Storage Subsequently, the XML structured article text is stored as a Wikisource item, which can be done utilizing WikiBots and converters that are already in place, see (Raspberry, Mietchen, and Taraborelli 2016).[4] [5] During this process, images from the article are stored as Wikimedia Commons objects, including technical metadata, accompanying license conditions, and references to their respective article on Wikisource. After that, for all items – articles as well as images – metadata sets in Wikidata are created.

Metadata extraction from text In a separated process, the article text is then processed using text mining methods, in order to enrich these Wikidata items. First the captions of the images are identified as well as the references to the image and the text surrounding the reference. Subsequently named entities and keywords (including multiterm keywords) are extracted from the text snippets related to an image. Finally, all found keywords and named entities are disambiguated and mapped onto standardized terms from the Wikidata thesaurus. The mapping is realized by using various synonym dictionaries and statistical methods like distributional similarity.

Image Classification A rough classification of images into categories like b/w, color, photograph or line drawing will be performed using OpenCV.

Indexing Now they are prepared to be ingested by indexing for Wikipedia's own search, as well as third party indexers, e.g. for inclusion into academic search engines, library discovery services etc.

The data can be corrected via crowdsourcing methods framed by Wikidata.

[3] http://oaspa.org/

[4] https://github.com/wpoa

[5] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Open_Access/

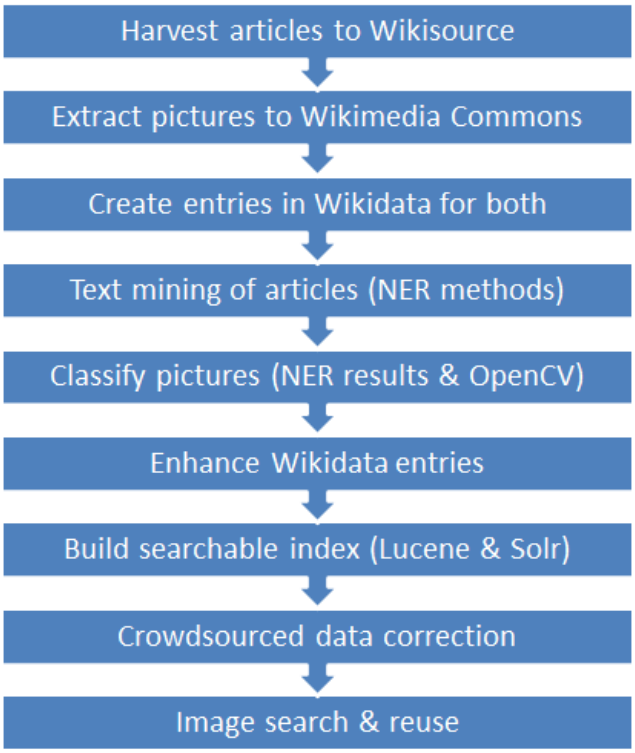

Figure 1: Processing of images and raw article text using Wikimedia infrastructure

Expected outcome

The main outcome of the suggested approach will be the provision of an information service which improves the retrieval of and the access to freely licensed images and figures from the fields of engineering and technology. It provides a reusable workflow and example for automatic collection, identification and availability of pictures from open access literature.

References

Blümel, I.; Cartellieri, S.; Heller, L.; and Wartena, C. 2014. Entwicklung eines Verfahrens zur automatischen Sammlung, Erschließung und Bereitstellung multimedialer Open-Access-Objekte mittels der Infrastruktur von Wikimedia Commons und Wikidata.

Raspberry, L.; Mietchen, D.; and Taraborelli, D. 2016. Wikipedia:WikiProject Open Access.